

# TELECOM CHURN PREDICTION



**Submitted by:**

*Varun Duggal*

*Priyanka Amit Sadhwani*

# INTRODUCTION

**Churn prediction** is one of the most popular Big Data use cases in business. It consists of detecting customers who are likely to cancel a subscription to a service.

**Churn** is a problem for telecom companies because it is more expensive to acquire a new customer than to keep your existing one from leaving.

Wireless companies today measure voluntary churn by a monthly figure, such as 1.9 percent or 2.1 percent.



# PROJECT OBJECTIVE


- To predict Customer Churn.
- Highlighting the main variables/factors influencing Customer Churn.
- Use various ML algorithms to build prediction models, evaluate the accuracy and performance of these models.
- Finding out the best model for our business case & providing executive summary.

There are many ways: better products, better delivery methods, lower prices, building satisfactory customer relationships, better marketing and, above all, successful customer communications.

# DATASET DESCRIPTION

- Source dataset is in csv format.
- Dataset contains 99999 rows and 226 columns
- Missing values found for the columns and column>30% missing values dropped.
- High value customers are variable of interest as 80% of revenue comes from them

# METHODOLOGIES

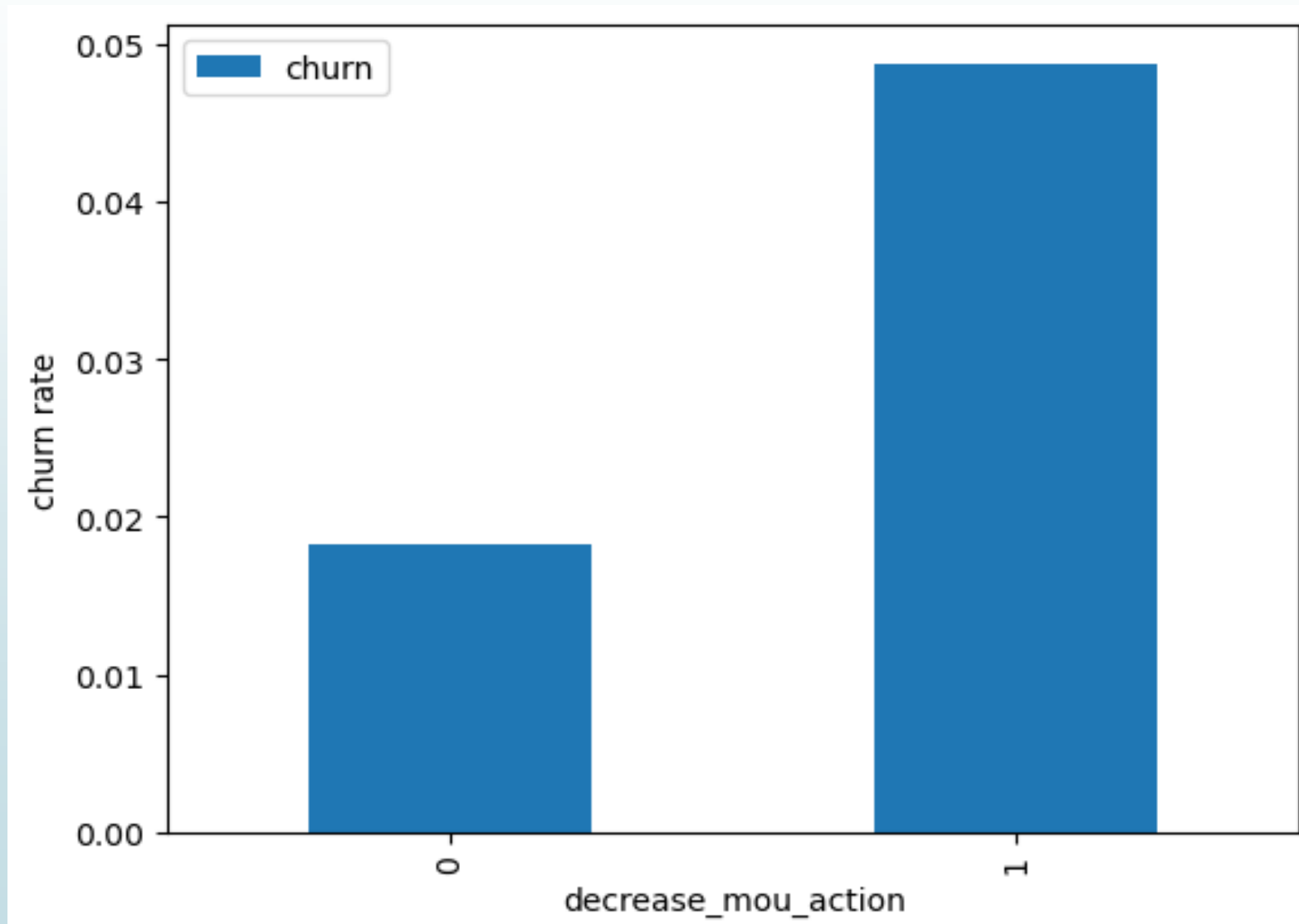
- 
- Methodology followed simple basic steps:
    - Calculation of missing values
    - Deletion of date columns
    - Filtering of high level customers
    - Handling of missing values
    - Tagging of churned customers based on fourth month
    - Checking of churn percentage
    - Outlier treatment
    - Feature derivation
    - EDA with Univariate and Bivariate analysis
    - Splitting of data into train test split(80:20)
    - Handling data imbalance
    - Feature scaling to remove data inequality and for standardization



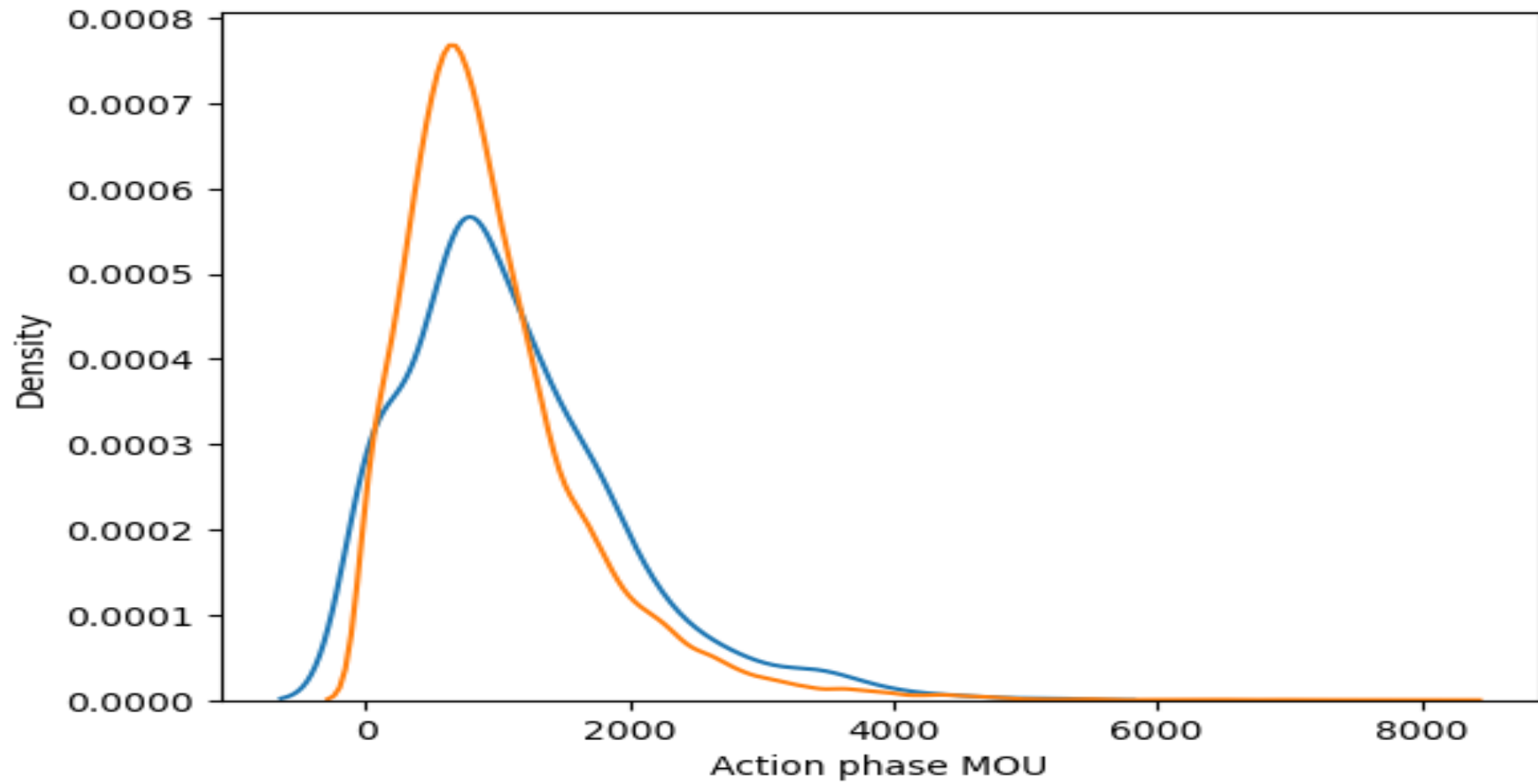
# METHODOLOGIES (continued)

- PCA and Logistic Regression
- Building model with optimal hyperparameters
- Prediction on test and train set
- Random forest with PCA along with hyperparameter tuning
- Feature selection using RFE
- Analysis of various models
- Plotting of ROC Curve

# Churn rate when customer decreased her/his recharge amount in action month

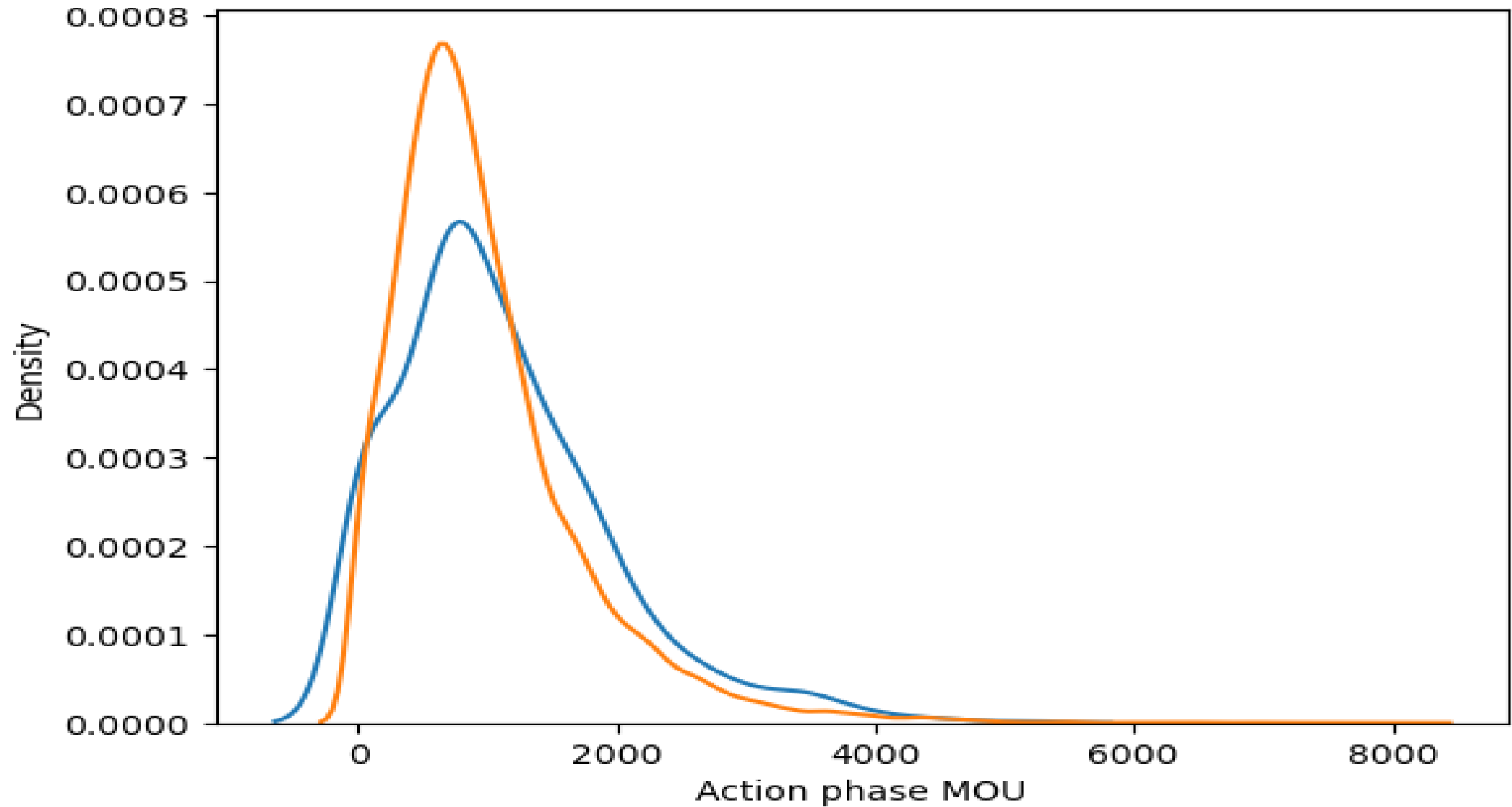


# ARPU

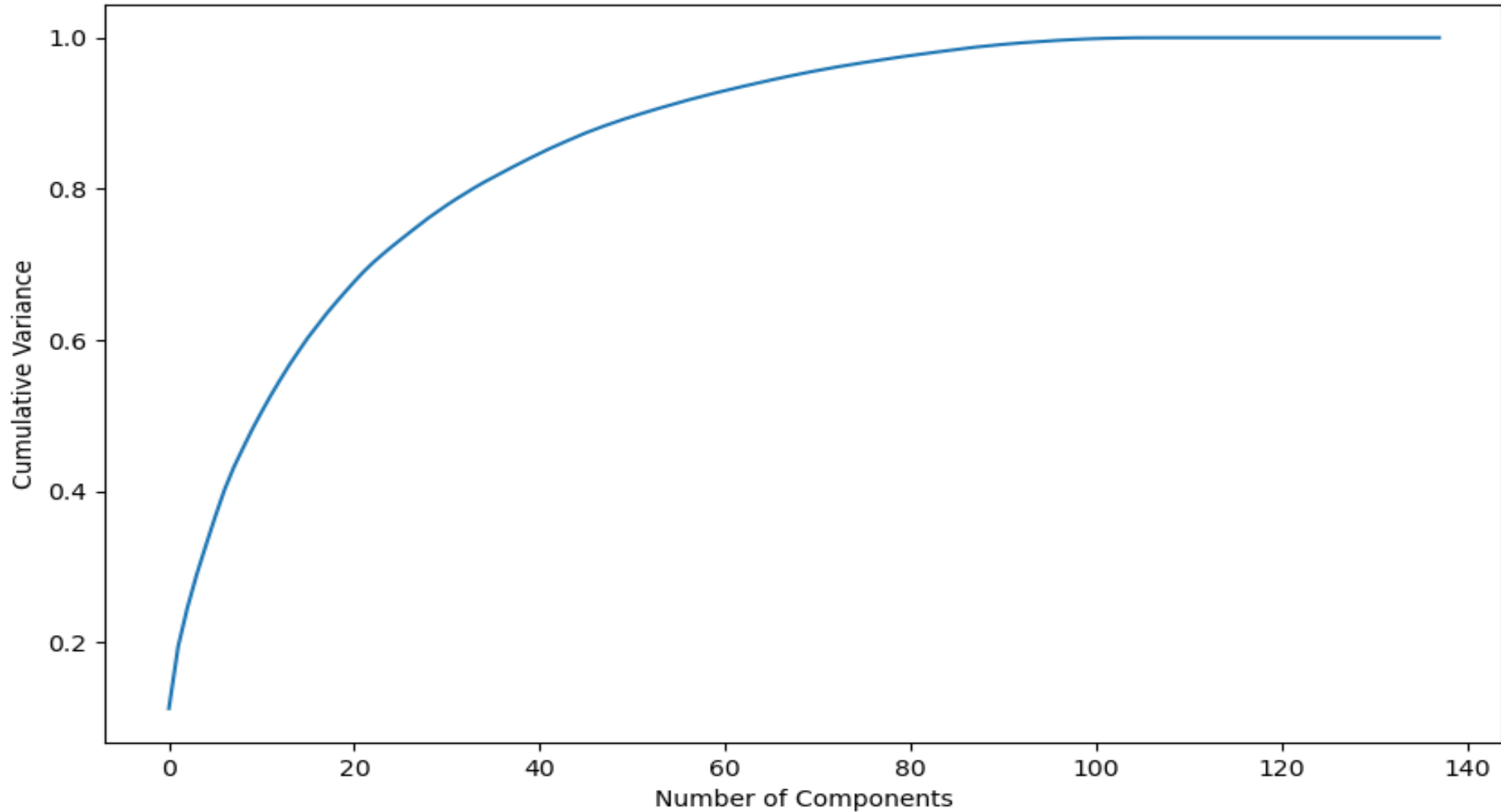




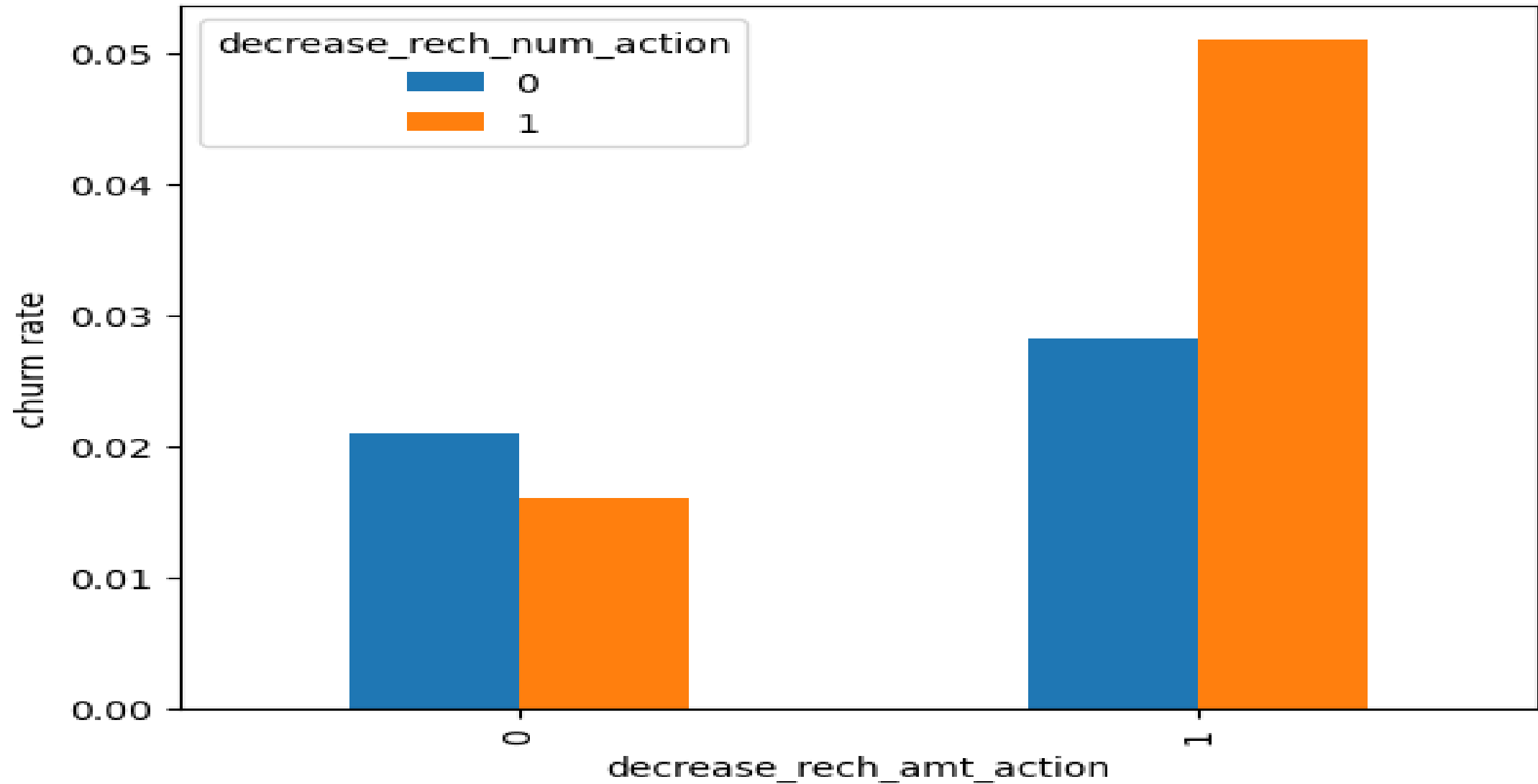
## MOU in action Phase



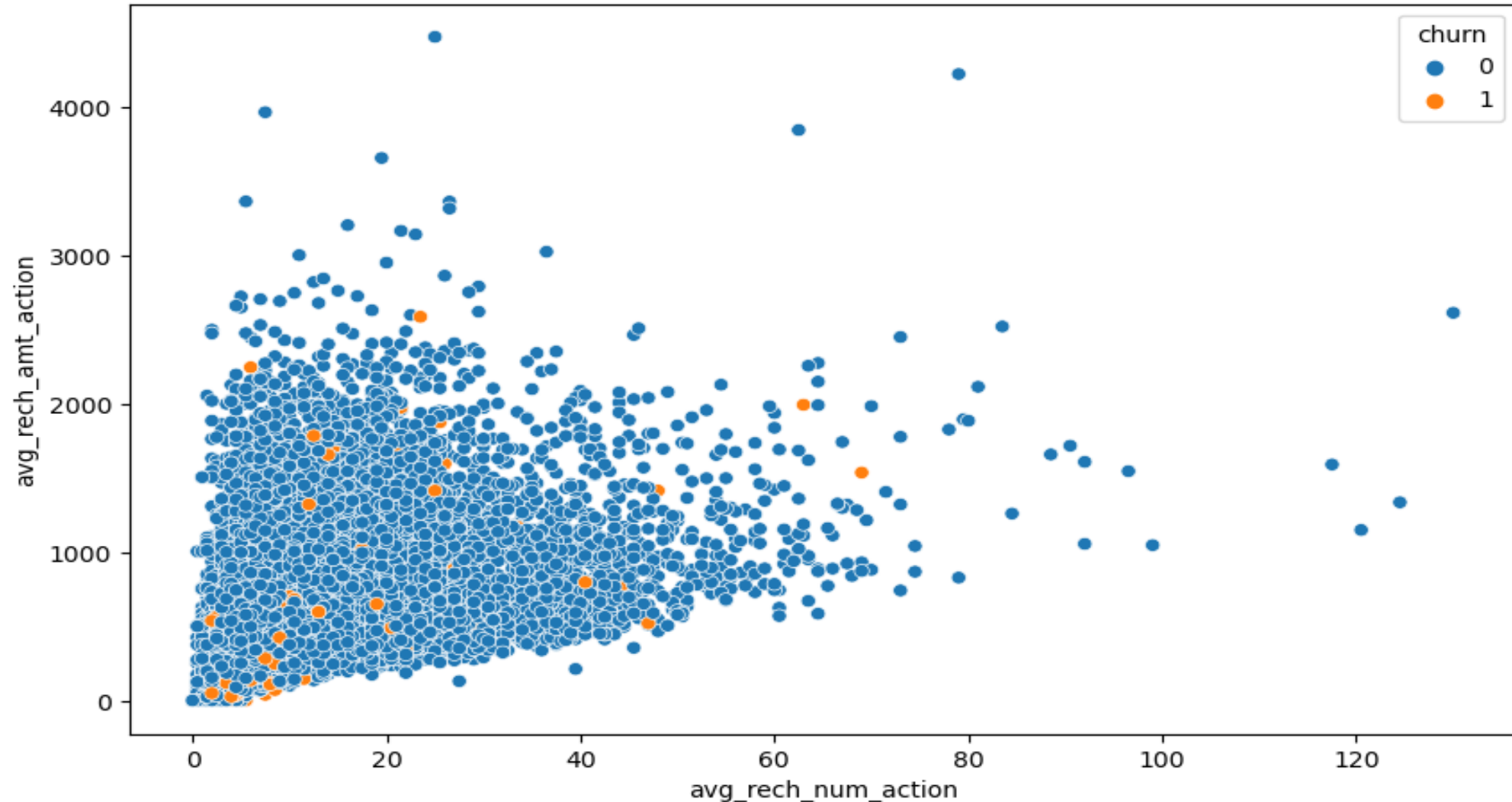
# PCA Graph for selection of components



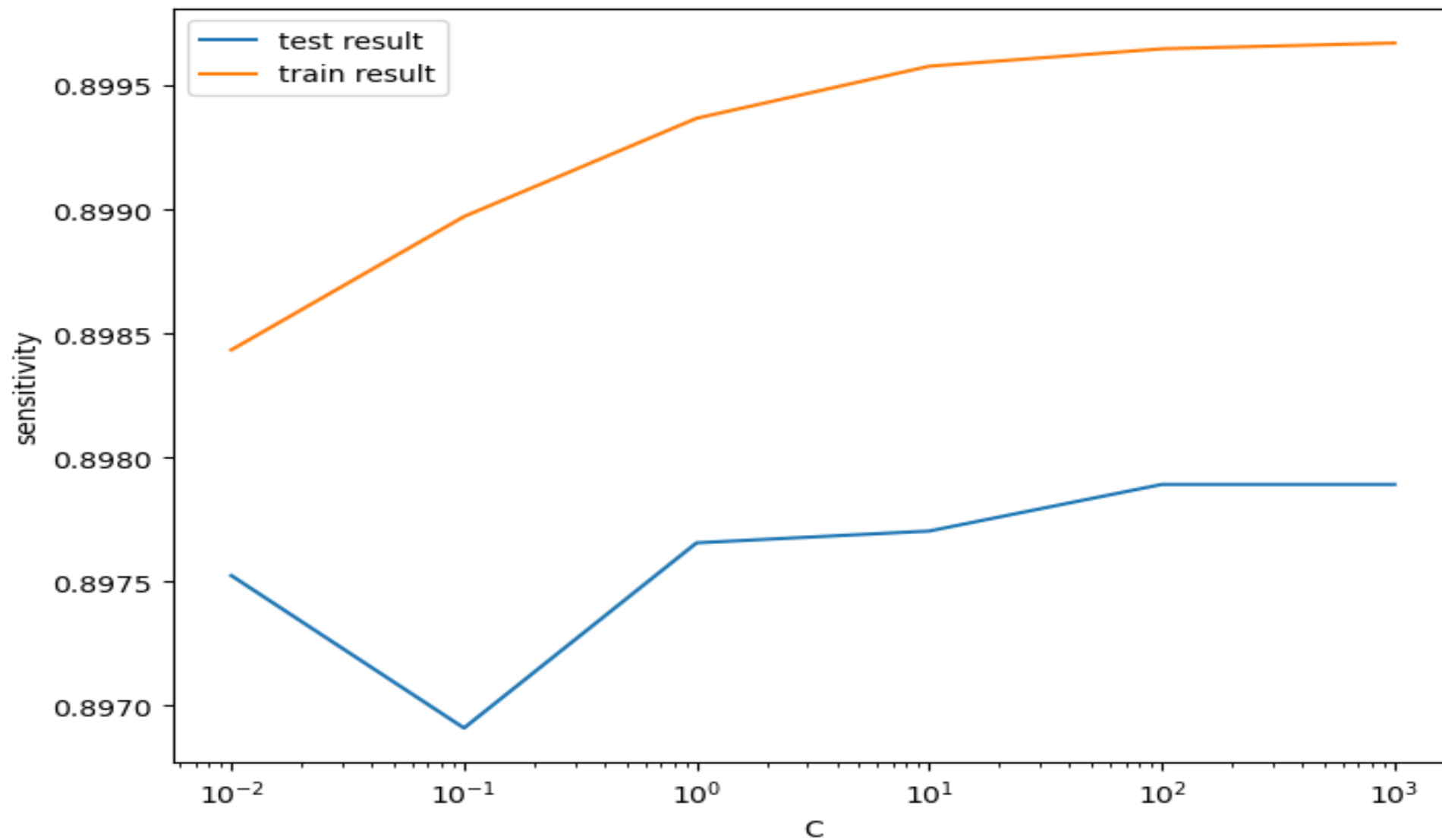
# Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase



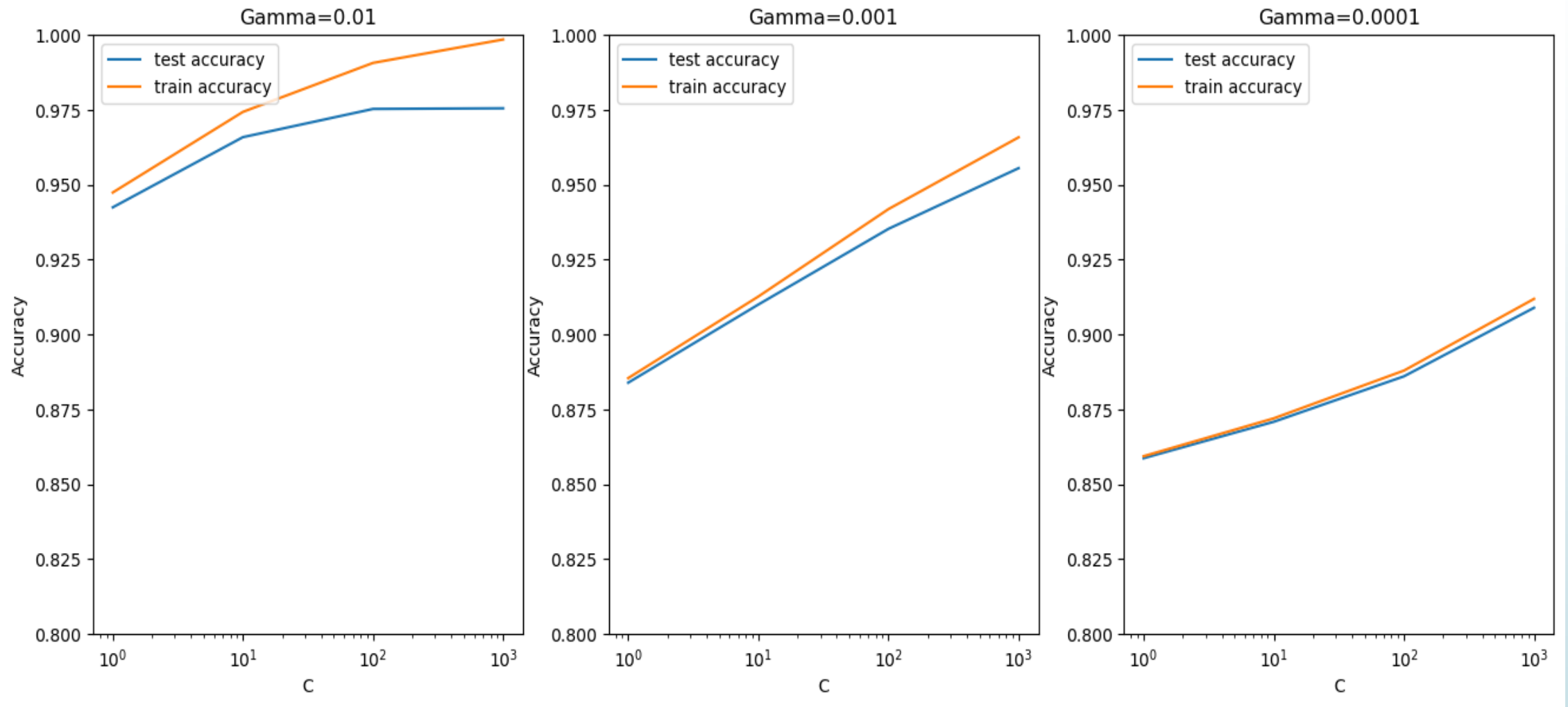
# Analysis of recharge amount and number of recharge in action month



# C vs Validation Score



# Accuracy with Gamma values



The best test score is 0.97 corresponding to hyperparameters {'C': 1000, 'gamma': 0.01}

# Prediction on train and test set

## ➤ Train set

- Accuracy = 0.89
- Sensitivity = 0.92
- Specificity = 0.85

## ➤ Test set

- Accuracy = 0.85
- Sensitivity = 0.81
- Specificity = 0.85

# Decision Tree

- Best sensitivity:- 0.9007
- DecisionTreeClassifier (max\_depth=10, min\_samples\_leaf=50, min\_samples\_split=100)

## **Model summary**

- Train set
  - Accuracy = 0.90
  - Sensitivity = 0.91
  - Specificity = 0.88
- Test set
  - Accuracy = 0.86
  - Sensitivity = 0.70
  - Specificity = 0.87



# Recommendations

## **Conclusion with no PCA**

- Logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA. So, we can go for the more simplistic model such as logistic regression with PCA.
- LR explains the important predictor variables as well as the significance of each variable.
- The model also helps us to identify the variables which should be act upon for making the decision of the to be churned customers. Hence, the model is more relevant in terms of explaining to the business.

# Best Predictors from LR Model

<u>Variables</u>	<u>Coefficients</u>
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

Top variables *have negative coefficients*, it means, the variables are inversely correlated with the churn probability.

## **For example:**

If the local incoming minutes of usage (loc\_ic\_mou\_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

# Final Model Results

## Final model summary

- Train set
    - Accuracy = 0.84
    - Sensitivity = 0.81
    - Specificity = 0.83
  - Test set
    - Accuracy = 0.78
    - Sensitivity = 0.82
    - Specificity = 0.78
- Overall, the model is performing well in the test set, *what it had learnt from the train set.*

# Final Business Recommendations

- Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
- Target the customers, whose outgoing others charge in July and incoming others on August are less.
- Customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
- Customers, whose monthly 3G recharge in August is more, are likely to be churned.
- Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
- Customers decreasing monthly 2g usage for August are most probable to churn.
- Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- roam\_og\_mou\_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.



THANK YOU